Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

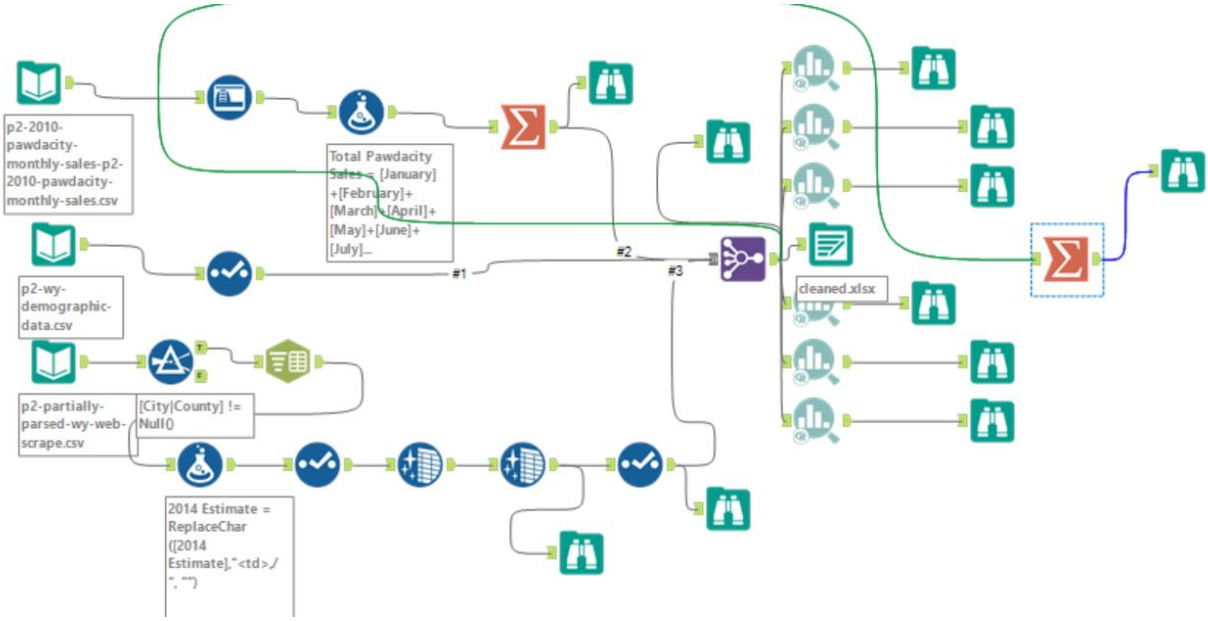
Key Decisions:

Answer these questions

1. What decisions needs to be made?
   Recommend the city for Pawdacity's newest store.

2. What data is needed to inform those decisions?
   Past Sales Data, Demographic Data of the cities, Population data, and the sales of competitor stores. These data are used to predict the yearly sales. **Predicted yearly sales** data used to answer the decision making.

Step 2: Building the Training Set



Column	Sum	Average
Census Population	*213,862*	19442
Total Pawdacity Sales	*3,773,304*	343027.6
Households with Under 18	*34,064*	3096.73
Land Area	*33,071*	3006.49
Population Density	*63*	5.71
Total Families	*62,653*	343027.6

Step 3: Dealing with Outliers

Answer these questions

Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

After doing analysis in excel (Pictures below). There are 4 outliers in the training set they are Casper, Cheyenne, Gillette, Rock Springs.

I have chosen to remove Cheyenne because:

a. It falls into outlier category in 5 fields (all except Land Area)
b. It varies significantly from the IQR than other outlier cities
c. Imputing the data for this record would adversely impact the model, thus removing the record is more appropriate

City	Land Area	Households with Under 18	Population Density	Total Families	Total Pawdacity Sales	2010 Census
Buffalo	3115.5075	746	1.55	1819.5	185328	4585
Casper	3894.3091	7788	11.16	8756.32	317736	35316
Cheyenne	1500.1784	7158	20.34	14612.64	917892	59466
Cody	2998.95696	1403	1.82	3515.62	218376	9520
Douglas	1829.4651	832	1.46	1744.08	208008	6120
Evanston	999.4971	1486	4.95	2712.64	283824	12359
Gillette	2748.8529	4052	5.8	7189.43	543132	29087
Powell	2673.57455	1251	1.62	3134.18	233928	6314
Riverton	4796.859815	2680	2.34	5556.49	303264	10615
Rock Springs	6620.201916	4022	2.78	7572.18	253584	23036
Sheridan	1893.977048	2646	8.98	6039.71	308232	17444
Average	3006.489126	3096.727273	5.709090909	5695.708182	343027.6364	19442
Q1	1845.593087	1423.75	1.95	3229.54	238842	9793.75
Q3	3448.432674	3015.5	5.1625	6327.14	304506	18842
IQR	1602.839587	1591.75	3.2125	3097.6	65664	9048.25
Upper Limit	5852.692054	5403.125	9.98125	10973.54	403002	32414.375
Lower Limit	-558.6662931	-963.875	-2.86875	-1416.86	140346	-3778.625

Scatter Plots (2 of 5, to emphasize the justification of Cheyenne as Outlier):